Presentation on the Occasion of Receiving the ADHO Antonio Zampolli Prize on Behalf of the TEI Community

NANCY IDE

DEPARTMENT OF COMPUTER SCIENCE

VASSAR COLLEGE

POUGHKEEPSIE, NEW YORK USA

Part One: Context and Rationale

- How was the TEI started?
- Why was the TEI started?
- What problems was it trying to solve?

Humanities Computing in the Mid-80s

- Vocabulary, authorship, stylistic studies
  - Concordance-based:
    - Words, word patterns, combinations
  - Basic statistics:
    - Sorted frequencies of letters, words, phrases
    - Type-token statistics
    - Ranking collocates by strength of association
    - Vocabulary distributions over a text
    - Etc.

Software

- Concordancing, frequency lists, etc.
  - Oxford Concordance Program (and MicroOCP)
  - University of Toronto’s Text Analysis Computing Tools (TACT)
  - WordCruncher
  - . . .
- Input formats varied dramatically!
  - Differed from program to program, project to project

Encoding practices

- Scores of schemes developed in 60s, 70s, 80s for
  - Representing special characters
  - Encoding logical divisions of text
  - Representing analytic or interpretive information
  - Reducing text critical apparatus to a linear sequence

The Problem

- Substantial editing required to use a text encoded for one program or purpose with another
    - . . . if even possible
- It was a mess!
    - Or, as one attendee at the Poughkeepsie meeting in 1987 put it, “chaos”

Examples

- Citation formats
  - Include an abbreviated form of a citation reference at the beginning of each line

```
VirAen01001arma virumque cano, Troiae qui primus ab oris
```

- COCOA format
  - Enclose references in angle brackets, embed in text

```
<W Shakespeare>
<T Merchant of Venice>
<A 2>
<S 6>
<C Graziano>
This is the penthouse under which Lorenzo
. . .
```

W = writer
T = title
A = act
S = scene
C = speaker
L = line number (or program can count if true to text)

Heavily influenced by hardware and software restrictions of the time

Early Attempts

- 1967
  - Martin Kay argues for a “standard code in which any text received from an outside source can be assumed to be”
- 1970s and early 80s
  - Discussion of a standard at various meetings of humanities scholars (San Diego 1977, Pisa 1980)
  - No consensus on how, or even whether, a standard should be developed

1986-7

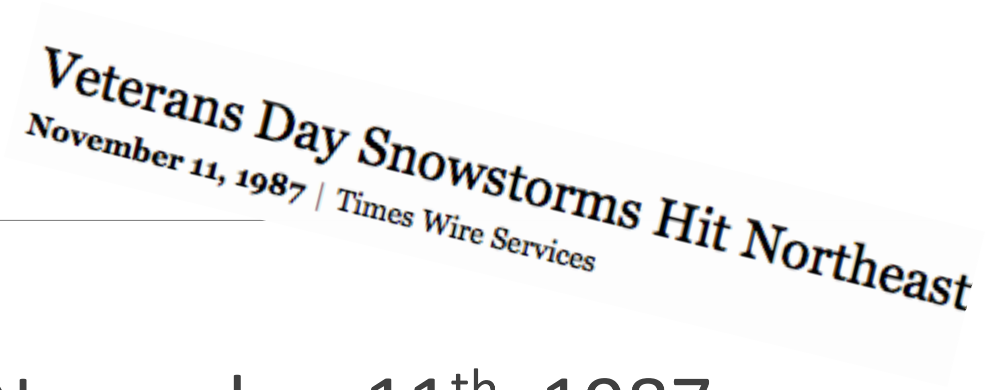
- Still plenty of discussion of need for a standard
- 1987 ICCH conference
  - Nancy Ide and Michael Sperberg-McQueen convince U.S. National Endowment for the Humanities representative **Helen Aguerra** to fund a workshop organized by the **Association for Computers and the Humanities (ACH)**, to bring together the relevant people to determine how and if such a standard would be possible

The Rest is History

- November 11-12, 1987: NEH Workshop at Vassar (Poughkeepsie)
  - Thirty-two people from around the world attended
    - Representatives of text archives, humanities computing centers, professional organizations
      - Organizations: ACH, ALLC originally
      - Antonio Zampolli involved Don Walker of the Association for Computational Linguistics (ACL)

THE HILLARD ROOM
NO SMOKING

Glitch in the Plan



Veterans Day Snowstorms Hit Northeast

November 11, 1987 | Times Wire Services

- Snowstorm in NY on November 11th, 1987
  - Travel from NY airports to Poughkeepsie very tricky!
  - Zampolli convinced a van driver to bring a group of participants stranded at JFK Airport to Poughkeepsie, despite the snow

The Workshop

- Two days of intense discussion led to agreement on
  - Need for common practice
  - Set of basic principles to guide the development of guidelines for encoding and exchange of literary and linguistic data

The “Poughkeepsie Principles”

Motivating Background (1)

- **Hardware constraints** had a huge impact on encoding choices
- **Existing software could be difficult** for the non-computer scientist to install and use
- Accompanying **documentation and metadata hard to specify** in a readily available, consistent way
- Notion of **separating prescriptive markup (how it looks) from descriptive markup (what it is)** was brand new (1986)
- Specter of the argument that it would be **impossible to define a single standard** that suits everyone

Resulting Principles

- Provide **descriptive** rather than **prescriptive** markup
- Provide means for **in-document metadata**
- Focus on **representing the required/desired information**, not on software requirements
- Define a scheme that is **hardware-, software-, and application independent**

Resulting Principles

The scheme should

- Be **simple, clear, and concrete**
- Be **easy for researchers to use** without special-purpose software

But at the same time:

- Allow for the **rigorous definition** and **efficient processing** of texts

Motivating Background (2)

The Poughkeepsie Principles everywhere reflect concern of archive representatives:

- **Requirement for retrospective conversion** of existing encoded texts
- **Loss of investment** in local expertise, software, and systems

Resulting Principle

The guidelines are intended to **suggest** principles for the encoding of texts in the format

Guidelines, not a standard!

Resulting Principle

The guidelines are intended to provide a standard format for **data interchange** in humanities research

- No requirement for conformance locally
- TEI scheme will serve as a “pivot” format
  - Only transduction of local format to and from TEI scheme (vs. *n*-way transduction among schemes)

Resulting Principle

The guidelines should define a **recommended** syntax for the format

- No final decision in Poughkeepsie on the exact syntax
  - SGML was promoted by many, but not unanimously accepted

Resulting Principle

The guidelines should include a **minimal** set of conventions for encoding new texts in the format

- **No requirements** will be made for the addition of information not already coded in the texts
- Newly-encoded texts should include **descriptive and bibliographic information**, and **information about the encoding itself**
- A **recommendation**
- Include means to **extend** the scheme

The Single Unfulfilled Principle

- The TEI project originally intended to **define a metalanguage for the description of text-encoding schemes**, and describe the new format and representative existing schemes in that metalanguage
  - Abandoned this goal because
    - Anxiety over translation of existing schemes subsided as TEI took shape
    - SGML gained far wider acceptance after the Vassar meeting
    - Volume of new texts being encoded shifted the balance of concern away from converting legacy data

Other Principles

- **Polytheoricity**
  - Little (or no) unanimity concerning relevant features to encode
  - Balancing act:
    - **Preserve intellectual autonomy** of researchers, but at the same time provide enough guidance to **avoid pointless variations** in encoding
  - Solution:
    - Specific DTD, but also **alternative means to encode the same thing** when felt necessary

Management

- Entrusted to a **Steering Committee** with representatives of three supporting bodies:
  - Association for Computers and the Humanities (ACH),
  - Association for Literary and Linguistic Computing (ALLC)
  - Association for Computational Linguistics (ACL)
  - This group raised over a million dollars in North America and Europe to support the work of TEI
  - Oversaw the development of the TEI Guidelines until 1996

Why Success This Time?

- **More known about encoding problems** and basic principles than in the past
- Included a **more robust representation** of key organizations and active research centers
- Recent **development of SGML** provided the right tool for a simple, flexible, and extensible encoding scheme

Reflection

- The size, scope, and influence of the TEI far exceeded what anyone at the Vassar meeting envisaged
- In retrospect, it is amazing to see how many foundational issues were addressed by the TEI
    - TEI as a “pivot” (interchange) format
    - Problem of polytheoreticity
    - Adherence to existing standards where possible
    - Requirement to include bibliographic information and description of encoding scheme

Many of these issues still operative in efforts to develop text representation standards

(Over to you, Michael)